VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHISC AND MULTIMEDIA

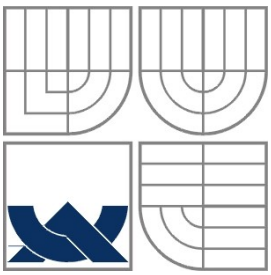
POROVNÁNÍ A PŘEVOD DATABÁZÍ SIGNÁLNÍCH DRAH

BAKALÁŘSKÁ PRÁCE
BACHELOR´S THESIS

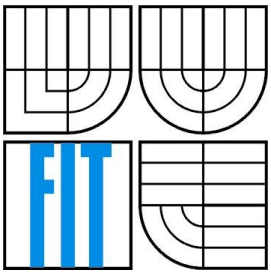
AUTOR PRÁCE
AUTHOR

PAVEL NĚMEČEK

BRNO 2008



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

POROVNÁNÍ A PŘEVOD DATABÁZÍ SIGNÁLNÍCH DRAH

COMPARISON AND TRANSFORMATIONS OF PATHWAY DATABASES

BAKALÁŘSKÁ PRÁCE
BACHELOR´S THESIS

AUTOR PRÁCE
AUTHOR

PAVEL NĚMEČEK

VEDOUCÍ PRÁCE
SUPERVISOR

Doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2008

Abstrakt

Tato práce se zabývá porovnáním a převodem databází signálních drah. Značná část je věnována databázím KEGG a BioCarta. V práci jsou rozebrány formáty poskytovaných dat. Při hledání překryvů drah mezi databázemi zkouším aplikovat shlukování, srovnání editační vzdálenosti názvů, porovnání genů podle čísla NCBI a nalezení ekvivalentních vazeb. Aplikace je implementována v jazyce C++ s použitím XML parseru. Výsledky jsou prezentovány ve formě textových výstupů, případně ve formě grafů zapsaných v jazyce DOT pro zpracování programem GraphViz.

Klíčová slova

Signální dráha, databáze KEGG, KGML, databáze BioCarta, PID (Pathway Interaction Database), Graphviz, Levenshteinova vzdálenost, NCI (National Cancer Institute), NCBI (National Center for Biotechnology Information)

Abstract

The main topic of this work are comparison and transformations of pathway databases. Large part is dedicated to databases KEGG and BioCarta. In work are described formats of provided data. In process of searching for overlapping pathways between databases are used clustering, edit distance, NCBI gene numbers and equivalent relationships comparison. The system was written in C++ and using XML parser. Results are text outputs, eventually files in DOT language, which can be executed with GraphViz to generate graphs.

Keywords

Signaling pathway, database KEGG, KGML, database BioCarta, PID (Pathway Interaction Database), Graphviz, Levenshtein distance, NCI (National Cancer Institute), NCBI (National Center for Biotechnology Information)

Citace

Pavel Němeček: Porovnání a převod databází signálních drah, bakalářská práce, Brno, FIT VUT v Brně, 2008

Porovnání a převod databází signálních drah

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením Doc. RNDr. Pavla Smrže, Ph.D. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Pavel Němeček
14. května 2008

Poděkování

Tímto děkuji vedoucímu práce za poskytnutí rad, kritiky a informačních zdrojů při vytváření programu a psaní této práce.

Obsah

1 Úvod

Vymezení funkce proteinu a biologické dráhy je významný problém v post-genomovém období. (Nyní v době po rozluštění lidského genomu.) Komunity, vědecké týmy a nadšenci se snaží nalézt způsob, jak odhadnout buněčné funkce z analýzy biologických dat vyskytujících se v databázích po celém světě. Tyto databáze jsou vytvářeny a spravovány organizacemi, univerzitami, ale také soukromými subjekty, které k těmto datům mnohdy poskytují informace navíc, ovšem za zakoupení licence. Správci databází často poskytují aplikační programové rozhraní k přístupu do jejich databáze a někdy také software k analýze těchto dat. Jelikož je dat velké množství je zpracování, prohledávání, analýza sekvencí, struktur a funkcí biologických makromolekul, tedy hlavně DNA a proteinů, bez počítačů takřka nemyslitelná. Proto se v poslední době velice rozvíjí oblast bioinformatiky. Tato oblast spojuje lidi z informačních technologií, kteří vědí jak přistupovat k databázím a jak zpřístupnit data, a biology, kteří tyto data vyhodnocují a dívají se na ně z pohledu biologického.

S vymezením funkce proteinu souvisí interakce mezi nimi, které lze například pozorovat při příjmu signálu buňkou (signální transdukce). Interakce hrají důležitou roli v mnoha biologických procesech a onemocněních. Více o příjmu signálu buňkou je v kapitole 2.1.

Dalším odvětvím bioinformatiky je zjišťování predikce chování buňky na základě aktivity jednotlivých genů (genové exprese), zejména v souvislosti s nádorovými onemocněními. Za tímto účelem jsou dnes nejpoužívanější biočipy (microarrays).

Výsledky při hledání nových biologických drah a zjištěných funkcích proteinů lze většinou nalézt na webových stránkách databází. Nové postupy při analýze dat je možné dohledat v některé z vědeckých databází jako je ACM, PubMed nebo Inderscience.

Cílem mé práce bylo porovnat obsah databází, vyhledat nejpravděpodobnější překryvy a jednotně reprezentovat dráhu. S tím souviselo i seznámení se signálními dráhami a nalezení vhodných metod, které by byly obecně použitelné.

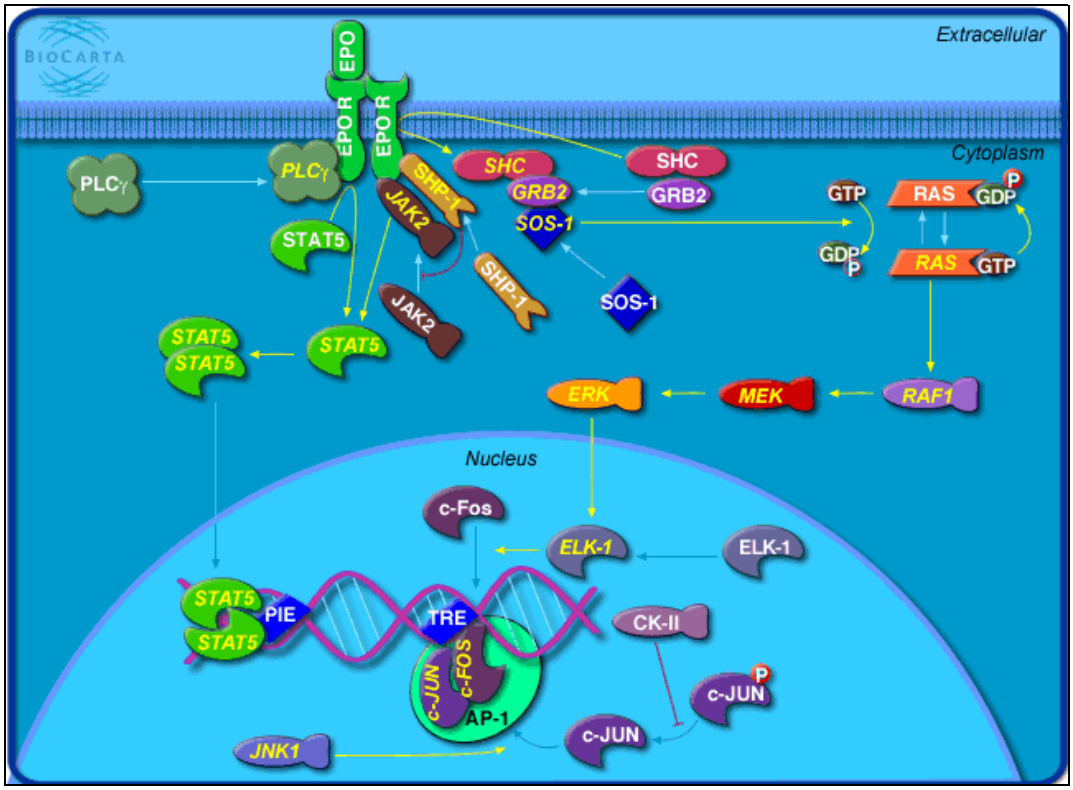
Následující kapitola se věnuje databázím biologických drah. Uvádím zde srovnání kvality databází BioCarta a KEGG a popisuji formáty dat, které zmíněné databáze používají. Dále uvádím způsob, jakým získané dráhy převádím na grafy, a problémy se kterými jsem se při převodu setkal. Ve třetí kapitole se zabývám metodami, které jsem použil při porovnávání drah. Čtvrtá kapitola se věnuje výsledkům, které program vypisuje, při různých typech porovnání. V této kapitole je také představen program Graphviz a jeho využití při zobrazovaní grafů. Poslední kapitola je závěr, který shrnuje výsledky práce a nastiňuje možnosti dalšího vývoje.

2 Databáze biologických drah

Existuje celá řada databází biologických drah, jejich přehled lze nalézt např. na stránkách Pathguide na adrese *http://www.pathguide.org/*. Pathguide obsahuje (v době psaní této práce) informace o 240 zdrojích biologických drah. Databáze jsou rozděleny podle obsahu do několika skupin, např. podle interakcí (protein-protein, protein-compound), podle povahy dráhy (metabolické, signální) atp. Pokud je to možné, je u databáze uvedena její dostupnost. Tedy zda je volně dostupná, dostupná jen pro akademické pracovníky, placená nebo prozatím nedostupná. Ve své práci se zabývám dvěma databázemi a to databází BioCarta a KEGG. Obě jsou volně dostupné a patří k nejpopulárnějším. KEGG z důvodu dobře udržované a ucelené databáze nejen signálních drah a BioCarta pro své názorné schémata drah. Mezi komerční databáze patří například MetaCore. Poskytuje množství rozšiřujících informací, které u mnoha volně dostupných databází nenalezneme.

2.1 Co je signální dráha

Signál v molekulární biologii mnohobuněčných organismů je informace vysílaná od jedné buňky ke druhé. Signální dráhy jsou sekvence událostí, které umožňují buňce přijmout signál a biologicky na něj reagovat [1].



Obrázek 2.1.1: Signální dráha EPO, zdroj BioCarta.

Informace se mezi buňkami šíří pomocí signálních molekul, které se váží se specifickými receptorovými molekulami na povrchu buňky, které umožňují signál přijmout a odpovědět na něj. Vazbou signální molekuly na receptor dojde ke změnám molekuly receptoru. Tato změna navodí sérii reakcí uvnitř buňky (signální transdukce), které vyústí v aktivaci určité buněčné činnosti [2]. Poznání těchto mechanismů může zlepšit diagnostiku i léčbu mnoha onemocnění. Postupně zkoumané a zpracovávané signální dráhy jsou ukládány do specializovaných databází. Pro ucelení představy, jak diagram signální dráhy vypadá, uvádím obrázek 2.1.1.

2.2 Databáze BioCarta

BioCarta byla založena v roce 2000 s cílem zaujmout vedoucí postavení ve vývoji, dodávání a distribuci zdrojů a vzorků pro biofarmaceutický a akademický výzkum.

Signální dráhy BioCarty ve formě jednoho XML souboru je možné stáhnout ze stránek NCI (National Cancer Institute) *http://pid.nci.nih.gov/PID/download.shtml.* NCI uchovává data v PID (Pathway Interaction Database), což je strukturovaná a spravovaná kolekce informací o známých biomolekulárních interakcích a klíčových buněčných procesech sestavených do signálních drah. Na uvedené stránce se nachází i popis struktury PID XML v jazyce DTD (Definice Typu Dokumentu). PID je společný projekt National Cancer Institute (NCI) a Nature Publishing Group (NPG) a je zaměřený na komunitu zkoumající rakovinu.

Databáze obsahuje biomolekulární interakce, které jsou v lidských buňkách známé nebo pravděpodobné. Databáze i redakční obsah jsou měsíčně aktualizovány. Formátu PID XML se budu více věnovat v kapitole 2.4.

2.3 Databáze KEGG

Projekt KEGG (Kyoto Encyclopedia of Genes and Genomes) byl iniciován japonským programem na výzkum lidského genomu. Od roku 1995 se snaží vyvíjet metody pro odvození vyššího systémového chování buňky a organismu z genomové informace. Databáze KEGG se skládá: ze signálních drah reprezentující rozpoznané molekulární interakce a reakční sítě (KEGG PATHWAY), z kolekce hierarchické klasifikace vztahů mezi biologickými systémy (KEGG BRITE), z informací o genech (KEGG GENES) a z chemických látek a reakcí, které jsou významné pro život (KEGG LIGAND).

KEGG poskytuje svoji databázi signálních drah ve formě XML souborů. XML reprezentace drah je aktualizována každé 2 až 3 dny. Každé signální dráze odpovídá jeden souboru. Na adrese *ftp://ftp.genome.jp/pub/kegg/xml* lze nalézt DTD, který popisuje strukturu XML souborů se signálními dráhami. KEGG tuto svoji speciální strukturu pro zápis drah pomocí XML nazývá KGML

(KEGG Markup Language), kterému se budu věnovat v kapitole 2.5. Na výše uvedené adrese lze nalézt také adresáře *organism*, *map* a *ko*. V adresáři *organism* jsou signální dráhy rozčleněny podle organismů, ve kterých se vyskytují. Ve své práci se zabývám jen lidskými dráhami (Homo Sapiens ---- zkratka KEGG hsa, podadresář *hsa*), aby je bylo možno porovnat s databází BioCarty, která obsahuje jen lidské dráhy. Adresář *map* obsahuje referenční dráhy. Tyto dráhy nemusí být společné všem organismům. Přibližně 2/3 z uvedených referenčních drah jsou společné i pro člověka. Poslední z adresářů je *ko*, který obsahuje signální dráhy zapsané s využitím čísel KO (KEGG Orthology). Čísla KO vznikají tak, že KEGG roztřídí všechny geny ve všech organismech do skupin funkčně identických genů (orthologs), a této skupině přiřadí jedno číslo KO [3]. Pomocí DBGET lze zjistit, které geny kterých organismů do které skupiny patří. DBGET je integrovaný databázový systém, kterému se dají přes webové rozhraní klást dotazy. Přehledné odpovědi jsou zobrazovány ve formě HTML stránek.

Mezi velice užitečné soubory z pohledu dalšího zpracování signálních drah KEGGu patří ty v adresáři na adrese *ftp://ftp.genome.jp/pub/kegg/genes/organisms/hsa/*.

Přehled některých souborů a jejich obsahu:

```
hsa_ko.list  -  "převodní tabulka" mezi KEGG ID číslem GENU a KO

hsa_ncbi-geneid.list  -  "převodní tabulka" mezi KEGG ID číslem GENU

                         a číslem genu v databázi NCBI

hsa_enzyme.list  -  "převodní tabulka" mezi KEGG ID číslem ENZYMU

                    a  KEGG ID číslem GENU

hsa_pathway.list  -  udává který gen se vyskytuje v které dráze
```

Důvodem pozastavení nad těmito soubory je to, že jejich použití v mé práci je podstatné při načítání drah z XML souborů a následné kontrole načteného obsahu.

2.4 Formát dat PID (Pathway Interaction Database)

PID obsahuje dva druhy dat.

1. NCI-Nature Curated data vytvořená editory Nature Publishing Group a zkontrolovaná odborníky. Molekuly jsou identifikovány UniProt (Universal Protein Resource) identifikátory. UniProt je komplexní katalog informací o proteinech dostupný na *http://www.expasy.uniprot.org/*. K interakcím jsou poznamenány evidenční kódy a reference, které lze využít k nalezení odborného článku v databázi PubMed (patří také pod NCI).

V těchto článcích jsou interakce popsány. Evidenční kód představuje kategorii článků. Reference ukazují přímo na určitý článek.

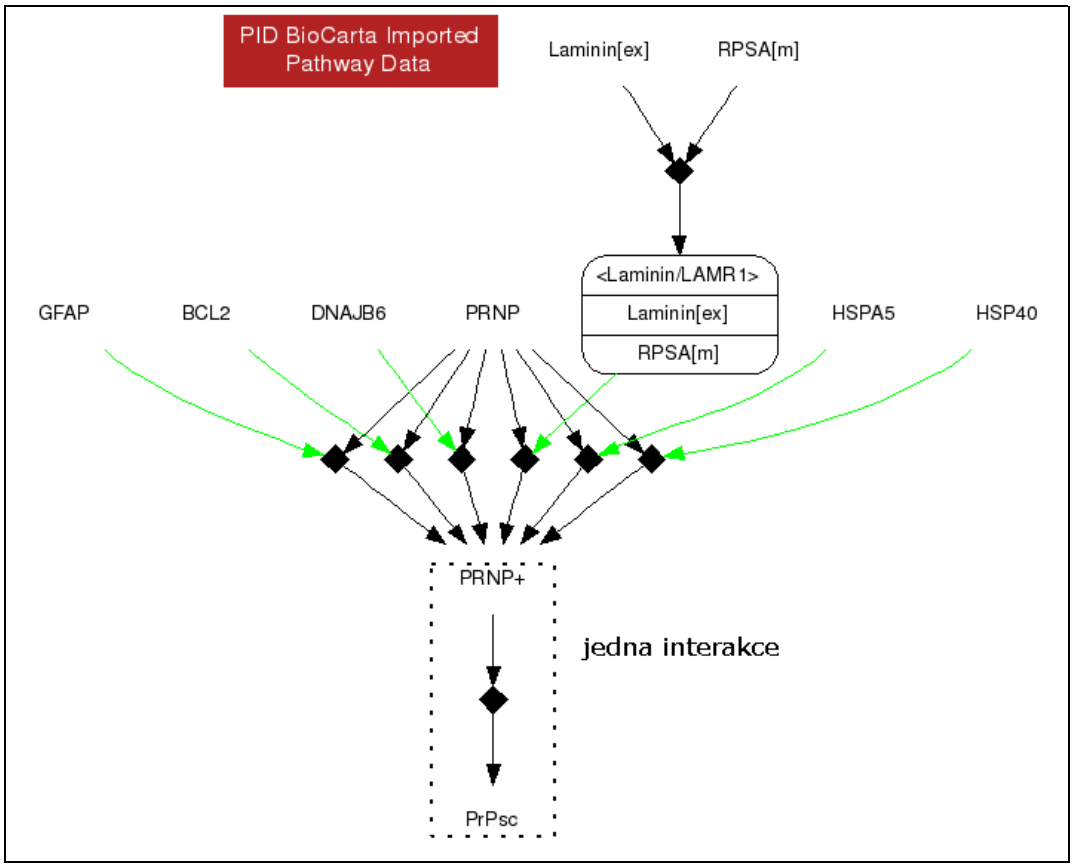
2. Data signálních drah z databáze BioCarty, která jsou součástí PID (Pathway Interaction Database) od roku 2004 a jsou importována bez odborné kontroly. Molekuly jsou identifikovány identifikátorem z databáze Entrez Gene (také patří pod NCI).

Ve zbytku kapitoly se zaměřím na popis struktury PID XML. V příkladech budu používat úseky PID XML databáze BioCarty.

Základní jednotkou informace v PID je molekulární interakce. Pro každou interakci databáze minimálně říká, které biomolekuly jsou v ní obsaženy (protein, RNA, komplexní biomolekuly, další složky), proces obsahující tyto biomolekuly (reakce, vazba, translokace, transkripce) a role každé biomolekuly v tomto procesu (vstup, produkt, agent, inhibitor). V signálních drahách jsou agenti a inhibitoři uváděni jako pozitivní (agenti) a negativní (inhibitoři) regulátoři a mohou působit na vstupní molekulu přímo či nepřímo.

V každé dráze jsou vstupní molekuly transformovány na výstupní molekuly. Tento proces může být podporován agenty nebo zpomalován inhibitory. Výstup z procesu může být vstupem, agentem nebo inhibitorem pro následující proces. Tím vzniká dráha navazujících interakcí.

Následující obrázek ukazuje grafickou reprezentaci dráhy BioCarty vykreslenou nástroji NCI.



Obrázek 2.4.1: Signální dráha Prion pathway ze stránek NCI

Uzly na obrázku 2.4.1 reprezentují biomolekuly nebo proces. Hrany popisují roli molekuly v procesu (input: černá šipka, output: černá šipka, agent: zelená šipka, inhibitor: červená šipka). Jedna interakce se skládá z jednoho procesu nebo molekulového uzlu, který je propojen vstupní a výstupní hranou se sousední molekulou nebo procesem. Na obrázku je interakce vyznačena tečkovaným obdélníkem.

Nyní již k samotnému PID XML. Kořenovým elementem je `<NCI_PID_XML>`, který obsahuje elementy `<ontology>` a `<model>`. Dceřiné elementy elementu `<ontology>` říkají, jaké typy hran, procesů a molekul se v drahách vyskytují. Ostatní informace z dceřiných elementů elementu `<ontology>` nezpracovávám. Element `<model>` obsahuje subelementy `<MoleculeList>`, `<InteractionList>` a `<PathwayList>`.

Element `<MoleculeList>` obsahuje všechny molekuly vyskytující se v drahách v souboru PID XML. Tyto molekuly mohou nabývat typů uvedených v subelementu elementu `<onthology>` s atributem `molecule-type`. Příklad molekuly:

```
<Molecule molecule_type="protein" id="101243">
    <Name name_type="LL" long_name_type="EntrezGene" value="5621" />
    <Name name_type="AS" long_name_type="alias" value="PrPc" />
    <Name name_type="OF" long_name_type="official symbol" value="PRNP" />
</Molecule>
```

Element `<InteractionList>` obsahuje všechny interakce, které se vyskytují v PID XML souboru s dráhami. Příklad interakce:

```
<Interaction interaction_type="modification" id="100869">
    <Source id="3">BioCarta</Source>
    <EvidenceList>
        <Evidence value="NIL">NIL</Evidence>
    </EvidenceList>
    <InteractionComponentList>
        <InteractionComponent role_type="input" molecule_idref="101243">
            <Label label_type="activity-state" value="active" />
        </InteractionComponent>
        <InteractionComponent role_type="output" molecule_idref="101248">
        </InteractionComponent>
    </InteractionComponentList>
</Interaction>
```

Tabulka 2.4.1: Zápis interakce v souboru PID XML vyznačené na obrázku 2.4.1

Element <PathwayList> sdružuje dráhy obsažené v PID XML souboru. Dráhy se skládají z interakcí, jak ukazuje následující příklad.

```
<Pathway id="100062" subnet="false">
    <Organism>Hs</Organism>
    <LongName>prion pathway</LongName>
    <ShortName>prionpathway</ShortName>
    <Source id="3">BioCarta</Source>
    <PathwayComponentList>
        <PathwayComponent interaction_idref="100868" />
        <PathwayComponent interaction_idref="100867" />
        <PathwayComponent interaction_idref="100862" />
        <PathwayComponent interaction_idref="100869" />
        <PathwayComponent interaction_idref="100866" />
        <PathwayComponent interaction_idref="100863" />
        <PathwayComponent interaction_idref="100865" />
        <PathwayComponent interaction_idref="100864" />
    </PathwayComponentList>
</Pathway>
```

Dokument PID XML je provázán pomocí id čísel. Tato čísla jednotlivých elementů nejsou jedinečná a tak se stává, že např. molekula má stejné identifikační číslo jako interakce. Na tento jev je nutné při zpracování dávat pozor.

2.5 Formát dat KEGG

Jak již bylo řečeno v kapitole 2.3, KEGG používá pro své signální dráhy KGML (KEGG Markup Language). V této kapitole daný formát přiblížím.

KGML je XML reprezentace signálních drah KEGG. KEGG používá KGML jako formát pro popis objektů v grafech, zejména pak při tvorbě signálních drah, které jsou manuálně kresleny a aktualizovány. To především znamená, že většina objektů v obrázku dráhy má jako své atributy informace o poloze (souřadnice x, y). KGML umožňuje automatické vykreslování drah a poskytuje možnost počítačové analýzy a modelování proteinových a chemických sítí. Více o proteinových a chemických sítích uvádím na konci této kapitoly.

KGML reprezentace signálních drah obsahuje kořenový element <pathway>, který přestavuje jeden objekt typu graf. Atributy dráhy specifikují její jméno, KEGG ID číslo, typ organizmu, ve kterém se vyskytuje, a odkaz na DBGET. Příklad elementu <pathway>:

```
<pathway name="path:hsa05060" org="hsa" number="05060"
    title="Prion diseases"
    image="http://www.genome.jp/kegg/pathway/hsa/hsa05060.gif"
    link="http://www.genome.jp/dbget-bin/show_pathway?hsa05060">
</pathway>
```

Element <pathway> obsahuje dva hlavní typy elementů. Prvním z nich jsou elementy <entry>, které představují uzly grafu. Podle hodnoty atributu type jimi mohou být enzymy, geny, KO (ortholog group), složky (compound), skupina (group - více genových produktů) a odkazy na jiné dráhy (map). Na obrázcích signálních drah KEGG jsou geny, KO, enzymy, skupiny a odkazy na jiné dráhy zobrazeny jako obdélníky s číslem či názvem entity, kterou zobrazují. Kolečky jsou znázorněny složky (compound). Ve většině případů jsou to odkazy do databáze, kde lze nalézt více informací o dané entitě. Element <entry> má ještě atribut name, který je KEGG ID číslem entity. Při vytváření uzlu zpracovávám číslo, jméno, typ a grafické jméno elementu <entry>.

```
<entry id="5" name="hsa:3309" type="gene">
    <graphics name="HSPA5" fgcolor="#000000" bgcolor="#FFFFFF"
        type="rectangle" x="217" y="120" width="45" height="17"/>
</entry>
```

Druhým typem objektů jsou vazby (hrany) mezi uzly grafu -- elementy <relation> a <reaction>. Element <relation> obsahuje atributy entry1 a entry2, které určují, mezi kterými uzly se vazba vytvoří. Typ vztahu určuje atribut type. Hodnoty, kterých může nabývat, jsou vypsány v tabulce níže. Tyto hodnoty jen upřesňují význam zaměřenosti vazby (v jakém kontextu mají být vazby chápány).

```
ECrel - vztah typu enzym-enzym

PPrel - vztah protein-protein, například vazba nebo modifikace

GErel - vztah typu genové exprese, indikační vazba transkripčního
        faktoru a výsledného genového produktu

PCrel - vztah protein-compound

maplink - odkaz na jinou dráhu
```

Pokud je uveden element <subtype>, potom jeho atributy value a name určují zaměřenost vazby. Hodnoty, kterých může atribut nabývat, jsou uvedeny v tabulce.

```
expression, activation     -->    exprese, aktivace
repression, inhibition     --|    potlačení, utlumení
indirect effect            ..>    nepřímý efekt bez molekulárních
                                  detailů
state change               ...    přechodové stádium
binding/association        ---    spojení
dissociation               -+-    rozklad
```

V následném porovnání drah mezi databázemi jsou navzájem ekvivalentní jen některé vazby. Více je popsáno v kapitole 3.3.

```
<relation entry1="1" entry2="11" type="PPrel">
    <subtype name="binding/association" value="---"/>
</relation>
```

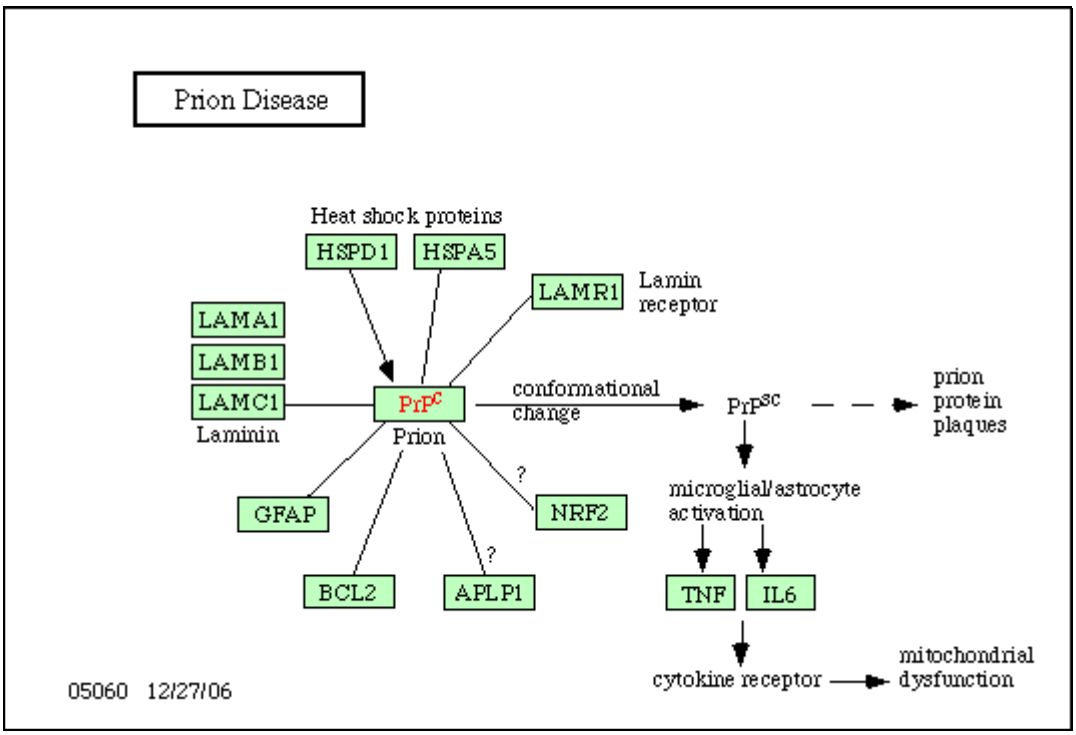
Tabulka 2.5.1: Příklad zápisu vztahu mezi dvěma geny v KGML

Element `<reaction>` specifikuje chemickou reakci mezi substrátem a produktem. Vztah znázorňuje šipka spojující dvě složky (compound). Element `<reaction>` má dva vnořené elementy: `<substrate>` a `<product>` (zdrojovou a cílovou složku vazby). Tento druh vazeb ve své práci nezpracovávám, protože jsem se při porovnávání zaměřil spíše na porovnání genů a vazeb mezi nimi než na zbývající složky dráhy. Pro úplnost ještě uvádím příklad vztahu `<reaction>`.

```
<reaction name="rn:R07771" type="irreversible">
    <substrate name="cpd:C16238"/>
    <product name="cpd:C16237"/>
</reaction>
```

Objekty grafu, které se skládají z elementů `<entry>` a `<relation>`, jsou nazývány proteinové sítě. Objekty grafu, které se skládají z elementů `<entry>` a `<reaction>` se nazývají chemické sítě. A na metabolické dráhy lze nahlížet jak na sítě proteinů (enzymů), tak jako na sítě chemických složek.

Nakonec uvádím grafickou reprezentaci dráhy, tak, jak ji na svých stránkách prezentuje KEGG.



Obrázek 2.5.1: Příklad signální dráhy hsa05060 ze stránek KEGGu

2.6 Srovnání kvality databází

V této kapitole se věnuji oběma databázím z pohledu kvality poskytnutých informací.

Začnu formou, jakou jsou poskytovány signální dráhy, tedy na úrovni souborů KGML (KEGG) nebo PID XML (BioCarta). NCI zpřístupňuje databázi signálních drah BioCarty ve formě jednoho souboru PID XML. Naproti tomu KEGG poskytuje každou dráhu v jednom souboru, což je velká výhoda při případné editaci dráhy a celkové orientaci v souboru. Je velký rozdíl editovat soubor s jednou dráhou a soubor, ve kterém se vyskytuje 254 drah (v době psaní této práce). S vyvinutím většího úsilí je možné získat dráhu BioCarty v jednom souboru na stránkách NCI (*http://pid.nci.nih.gov/PID/browse_pathways.shtml*). U této databáze mi chybí možnost přímého stažení souboru, tak jak je tomu u KEGG.

Další předností databáze KEGG je, že poskytuje soubory pro možnou částečnou kontrolu načteného obsahu z XML souborů. Jedná se především o soubor hsa_pathway.list zmíněný v kapitole 2.3. Lze tak překontrolovat chybějící geny a případně je pak podle informací na stránkách KEGGu doplnit. Co se týče kontroly obsahu je nutné dodat, že NCI na svých stránkách poskytuje jen čistě importovaná data BioCarty, která neprochází žádnou další kontrolou ze strany NCI, jak bylo popsáno v kapitole 2.4.

Možnou výhodou KEGGu je, že poskytuje signální dráhy mnoha organismů. Tuto přednost ovšem v práci nevyužívám, protože soubor BioCarty obsahuje jen signální dráhy lidské.

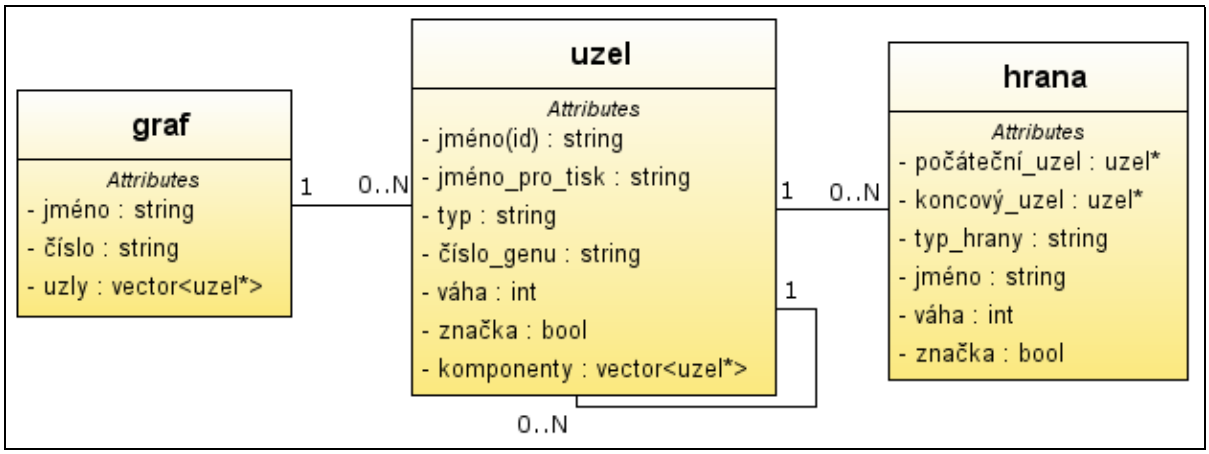
Z mé strany velmi dobré hodnocení databáze KEGG kazí fakt, že v některých souborech s lidskými dráhami neuvádí všechny vazby, tak jak jsou zobrazeny na obrázcích na oficiálních stránkách. Při práci s programem je tento jev patrný vždy, když je nápadně velké množství uzlů samostatných. Konkrétní příklad v době psaní práce je dráha Prion Disease (hsa05060). Vazby je nutné do XML doplnit např. z referenčního souboru map05060. Dalším nedostatkem KEGGu je, že v XML souborech nedostatečně využívá záznamy (`entry`) typu `group`, které shlukují více genů do jednoho uzlu. KEGG tvoří na obrázku uzel s více geny pomocí x-ových a y-ových souřadnic objektů. Pro přímé vykreslování pouhého obrázku to není špatný způsob, ale ke strojovému zpracování je nevhodný. V práci zpracovávám jen záznamy typu `group`. Naopak PID XML BioCarty je z pohledu vazeb a uzlů složených z více genů propracovanější a při testování jsem na žádný problém nenarazil.

2.7 Vlastní reprezentace pomocí grafu

Pro snadnější práci s dráhami jsem se rozhodl je reprezentovat jako neorientované grafy. Definice neorientovaného grafu [4]:

Neorientovaný graf je dvojice $G=\langle V, E\rangle$, kde V je neprázdná množina vrcholů (uzlů) a $E\in\{\{u, v\}|u, v\in V, u\neq v\}$ je množina dvouprvkových množin vrcholů (neorientovaných) hran.

Při implementaci grafu vycházím z definice, jen s tím rozdílem, že moje reprezentace dráhy může mít 0 až N uzlů a ty 0 až N hran. Obrázek ukazuje implementovanou strukturu.



U každého grafu je uvedeno:

```
jméno - název dráhy, př.: Prion diseases
číslo - identifikace dráhy, např.: hsa05060
seznam uzlů - uzly dráhy (geny, složky, makroproces, ...)
```

Pokud v souboru se signální dráhou není jméno uvedeno (KEGG: hsa05131), pak je nastaveno na "unknown pathway title".

Každý uzel obsahuje:

```
jméno - identifikační číslo v rámci dráhy
jméno pro tisk - většinou název genu nebo skupiny genů např.: BCL2
typ - typ uzlu (gen, složka, komplexní uzel, makroproces, ...)
číslo genu - NCI číslo genu, pokud je možné jej dohledat
váha - váha uzlu v grafu, zatím nepoužívané
značka - nastavením hodnoty si lze uzel v dráze poznačit
komponenty - uzel se může skládat z více uzlů a v komponentech jsou
             odkazy na tyto "poduzly"
hrany - seznam hran vycházejících z uzlu
```

Hrany vytvářím jen mezi uzly první úrovně. Pokud se vazba vyskytne mezi uzlem první úrovně a "poduzlem" nějakého uzlu první úrovně (např. při vytváření "group" u KEGGu), pak je tato vazba vyzvednuta na nejvyšší úroveň mezi uzly první úrovně.

Hrana obsahuje:

```
počáteční uzel - vždy ten samý uzel u kterého je hrana uvedena
                 v seznamu. Tato z prvního pohledu nadbytečná
                 informace mi usnadňuje práci při porovnávání hran.
koncový uzel - koncový uzel vazby
typ hrany - input, inhibitor, agent, -->, --|, ..>, ..., ---, -+-
jméno - např.: PPrel (protein-protein), více v kap. 2.5
váha - váha hrany v rámci dráhy, není využívána
značka - hranu si lze v grafu poznačit
```

Data ze souborů drah zpracovávám za pomocí XML parseru, který vytvořil Frank Vanden Berghen [5]. V prvním kroku jsou načteny samostatné uzly a hrany. Uzly jsou v KGML elementy `<entry>` a hrany elementy `<relation>`. V PID XML jsou uzly elementy `<Molecule>` a hrany elementy `<Interaction>`. Následně jsou hrany s uzly propojeny za pomoci id čísel XML elementů.

2.8 Problémy spojené s převodem

Mezi hlavní problémy spojené s převodem patří nevyváženost hloubky informace, které databáze poskytují. Tím je například myšleno to, že KEGG používá 6 hlavních typů vazeb a BioCarta jen 3. V kapitole 3.3 je uvedeno, jakým způsobem jsem si s tímto problémem poradil.

U některých molekul v souboru PID XML BioCarty není uvedeno číslo identifikátoru Entrez Gene, které je potřeba při porovnávání genů, i přesto, že jej lze v databázi Entrez dohledat. Pokud se toto číslo nenajde, porovnává se podle oficiálního názvu, příp. podle dalšího názvu (aliasu) genu. Obdobné pravidlo platí pro molekuly typu compound.

Další problémy souvisí se soubory databáze KEGG. Nalezené chyby jsou uvedené v tabulce:

```
hsa04140   - entry id "8"  K08340 není v převodní tabulce hsa_ko.list
hsa04330   - entry id "7"  K04497                    -||-
           - entry id "9"  KO6064                    -||-
hsa04340   - entry id "17" K06227                    -||-
hsa05120   - entry id "24" K01427                    -||-
hsa05217   - entry id "1"  K06227                    -||-

hsa04612   - v souboru chybí entry id "6"
           - neznámá vazba "==>"
hsa04660   - v souboru chybí entry id "50" a "51"
hsa04662   - v souboru chybí entry id "41"

hsa04115   - není uvedené číslo genu v databázi KEGG name="hsa:"
           - entry id "41"
hsa04730   - není uvedené číslo genu v databázi KEGG name="hsa:"
           - entry id "26"
```

3 Porovnání signálních drah

Mezi metody porovnání signálních drah jsem zařadil editační vzdálenost, shlukování, shodu genů v drahách, porovnání vazeb mezi společnými geny a doplněné srovnání o jiné složky než jsou geny. V následujících kapitolách budou jednotlivé metody rozebrány, budou nastíněny algoritmy metod, podle kterých jsem postupoval, a bude vysvětlen princip stanovení skóre. Skóre je číselné ohodnocení porovnání dvou entit, které je následně potřebné při seřazování výsledků.

3.1 Editační vzdálenost

Prvním z postupů, který jsem při porovnání signálních drah použil, je zjištění editační vzdálenosti mezi dvěma názvy drah. Existuje několik algoritmů k vypočítání této řetězcové metriky. Ve své práci používám Levenshteinův algoritmus, který je docela rychlý a přitom dostatečně reflektuje podobnost dvou řetězců.

Levenshteinova vzdálenost je dána minimálním počtem operací potřebných k transformaci jednoho řetězce na druhý. Mezi operace patří vložení, smazání nebo nahrazení jednoho znaku [6]. (Nahrazení znaku lze také chápat jako dvojici operací vložení a smazání).

př.: **krabice**

vrabec Levenshteinova vzdálenost je 3. Z pohledu změn na slově `vrabec` dojde k následujícímu: **v** nahrazeno **k**, **e** nahrazeno **i** a přidání **e** na konec slova

Při porovnávání názvů postupuji dvojím způsobem:

1. Z názvu dráhy vytvořím vektor slov, který poté porovnám s vektorem slov druhé dráhy. Porovnání probíhá každý s každým. Pokud je nalezeno odpovídající slovo (za odpovídající slova jsou prohlášena ta, jejichž editační vzdálenost je menší jak třetina délky prvního slova) je připočítáno skóre do hodnocení podobnosti názvů a slovo je z porovnávaného vektoru vyřazeno. Skóre odpovídající nalezené shodě dvou slov je:

$$2 * 100 / (\textit{počet slov nazvu } A + \textit{počet slov názvu } B)$$

Některá slova v názvech jsou považována za méně významná a jsou penalizována srážkou poloviny hodnoty skóre než při běžné shodě dvou slov. Mezi slova se sníženým významem patří: and, of, the, in a by.

2. Názvy drah jsou porovnány přímo jako dva řetězce. Tento způsob v porovnání s 1. lépe vyjadřuje skutečnou podobnost názvů v případě, že se srovnávají dva stejné názvy, v nichž se vyskytují méně významná slova (and, of, the, in a by), která jsou ohodnocena méně body.

Jako výsledná hodnota je vráceno lepší skóre z obou způsobů porovnání. Uživateli je zobrazeno desetinné číslo od 0 do 1, kdy 1 značí úplnou shodu názvů.

Při porovnávání drah je shodě v názvu přikládán maximální význam dvaceti procent z celkového hodnocení.

3.2 Shoda genů v drahách

Nejzajímavějším údajem při porovnávání dvou drah je počet shod genů. Geny jsou porovnávány na základě jejich čísla v databázi NCBI (National Center for Biotechnology Information). NCBI bylo založeno 1988. Je to americký zdroj pro informace o molekulární biologii. Vytváří veřejné databáze, vede výzkum ve výpočetní biologii, vytváří softwarové nástroje pro analýzu genomových dat a poskytuje biomedicínské informace [7]. Podrobné informace o jednotlivých genech je možné najít v databázi EntrezGene (spravovaná NCBI).

Samotné porovnání probíhá ve dvou fázích:

1. Nejprve jsou společné geny v obou dráhách (grafech) označeny, včetně těch, co se v dráze vyskytují vícekrát. To má význam při dalším porovnávání na shodu vazeb (hran) mezi společnými geny. Kdyby byl označen vždy jen jeden, tak bychom mohli přijít o některé společné vazby.
2. Následně jsou geny spočítány (započítán vždy jen jeden výskyt genu) a stanoveno skóre:

$$(\textit{počet společných genů} * 2.0) / (\textit{počet genů dráhy } A + \textit{počet genů dráhy } B)$$

Při porovnání drah je shodě genů v dráze přikládán největší význam, a to ve výši padesáti procent z celkového hodnocení.

3.3 Porovnání vazeb mezi společnými geny

Tato kapitola přímo navazuje na předcházející, ve které se označí v grafu (dráze) všechny společné geny. Pokud je úroveň hloubky porovnání vazeb nastavena na dva, jsou do obou drah mezi společné geny přidány "tranzitivní vazby", které spojují aktuální uzel s následníky jeho následníka. Nově přidané vazby jsou neorientované.

Porovnání probíhá následovně: Z grafů jsou extrahovány všechny vazby mezi společnými geny (i "tranzitivní") a je z nich vytvořena množina pro každou dráhu zvlášť. Následně jsou množiny

navzájem porovnány. To, zda jsi jsou vazby podobné, závisí na tom, zda spojují shodné geny (nebo aspoň jejich podmnožinu, jedná-li se o komplexní uzel) a zda mají shodnou samotnou vazbu. Tím je myšlena její orientace, případně typ. Naneštěstí databáze KEGG obsahuje 6 typů vazeb (nepočítáme-li typy molekulárních událostí) a databáze BioCarta 3. Aby bylo srovnání možné, je nejprve nutné najít odpovídající vazby. Udělal jsem následující transformaci:

- `KEGG: -->  odpovídá  BioCarta: input, agent`
- `KEGG: --|  odpovídá  BioCarta: inhibitor`
- `KEGG: ---  odpovídá  BioCarta: input, agent, output`
- `KEGG: -+- ... ..>  neodpovídá žádné vazbě z BioCarty`

"Tranzitivní hrana" je podobně jako hrana `---` kompatibilní se všemi ostatními hranami.

Shodné hrany jsou označeny, toho využiji při generování grafické reprezentace grafu pomocí nástroje Graphviz.

Skóre je spočítáno:

$$(\textit{počet společných vazeb} * 2.0)/(\textit{počet vazeb dráhy } A + \textit{počet vazeb dráhy } B)$$

V celkovém porovnání drah je porovnání na shodu vazeb přiřazena váha ve výši dvaceti procent z celkového skóre.

3.4 Doplněné srovnání o jiné složky než jsou geny

Toto srovnání vzniklo mimo jiné v reakci na to, že některé entity: geny, složky (compound), enzymy a RNA nemají ve svých atributech uvedenu hodnotu, která je jednoznačně identifikuje a podle které je možné je porovnat s jinými složkami v dalších databázích. Nedostatek těchto údajů se dá najít u databáze BioCarta.

Některé složky v databázi BioCarty obsahují identifikační číslo z databáze KEGG, což značí propojenost těchto databází a také nabízí jeden způsob, podle kterého složky porovnat. Druhý způsob, kterým se porovnává zbytek složek dráhy, který není nijak identifikován, je porovnání na shodu názvu entity. To může být její oficiální jméno nebo jen alias, podle toho, jaké informace jsem z databáze získal. Při porovnávání do grafů shodné složky neznačím.

Skóre je stanoveno :

$$(\textit{společných složek} * 2.0)/(\textit{počet složek } A + \textit{počet složek } B)$$

Tomuto porovnání je při celkovém výpočtu hodnocení podobnosti drah přisuzována nejmenší váha, činí deset procent z celkového hodnocení.

3.5 Shlukování

Shlukování je forma reprezentace dat, při které dochází za cenu jisté ztráty informace ke snížení objemu relevantních dat [8]. Aby bylo možné množiny shlukovat, musí entity v nich obsahovat stejné atributy (v našem případě to jsou čísla genů NCBI).

Shlukem je taková podmnožina $X \in O$, pro niž platí:

$$max(V(O_i, O_j)) < min(V(O_k, O_i)), O_i, O_j \in X, O_k \notin X \quad ,$$

kde O je množina objektů $O = O_1 \ldots O_n$ a V je míra vzdálenosti objektů.

Hierarchické metody shlukování se dají rozdělit na dvě hlavní skupiny. A to metody používající aglomerativní a rozdělovací algoritmy. Aglomerativní algoritmy postupují směrem zdola nahoru, kdy v konečném důsledku vznikne jeden shluk. Rozdělovací algoritmy postupují opačným směrem, kdy výchozí stav je jeden shluk. Hierarchické shlukování vytváří stromovou strukturu zvanou *dendrogram*. Proces vytváření shluků je většinou zastaven na určitém stupni hierarchie. V mém případě je proces zastaven až tehdy, když už není možné žádné dva shluky spojit, protože jejich podobnost je menší než stanovená hranice. Podobnost shluků je určována počtem shodných genů mezi shluky a zadává ji uživatel programu. Udává procentuální minimum společných genů v drahách. Při vytváření shluků postupuji podle následujících bodů:

1. Rozdělím množinu drah do shluků obsahujících po jedné dráze.
2. Vypočítám matici podobnosti.
3. Prohledám matici podobnosti, pokud naleznu dvojici, jejichž podobnost je vetší než stanovená hranice, spojím ji do jednoho shluku. To provedu pro všechny dvojice v aktuálním stupni hierarchie.
4. Pokud jsem provedl aspoň jedno spojení shluků, přejdu do bodu 2.

Matici podobnosti lze počítat jen pro horní či spodní trojúhelníkovou matici, protože je symetrické vůči hlavní diagonále. Příklad matice podobnosti vypadá následovně:

	s1	s2	s3	s4	s5
s1	x	0	0.5	0	0.7
s2	0	x	0.4	0	1
s3	0.5	0.4	x	0	0.2
s4	0	0	0	x	0
s5	0.7	1	0.2	0	x

Tabulka 3.5.1: Příklad matice podobnosti

Prvky s1 až s5 jsou shluky a hodnoty v matici udávají jejich podobnost (od 0 do 1), kdy 1 znamená nejvyšší podobnost.

Při shlukování drah uživatel nastavuje míru podobnosti. S touto hodnotou je nutné experimentovat, pak lze dojít k zajímavým výsledkům.

4 Výsledky

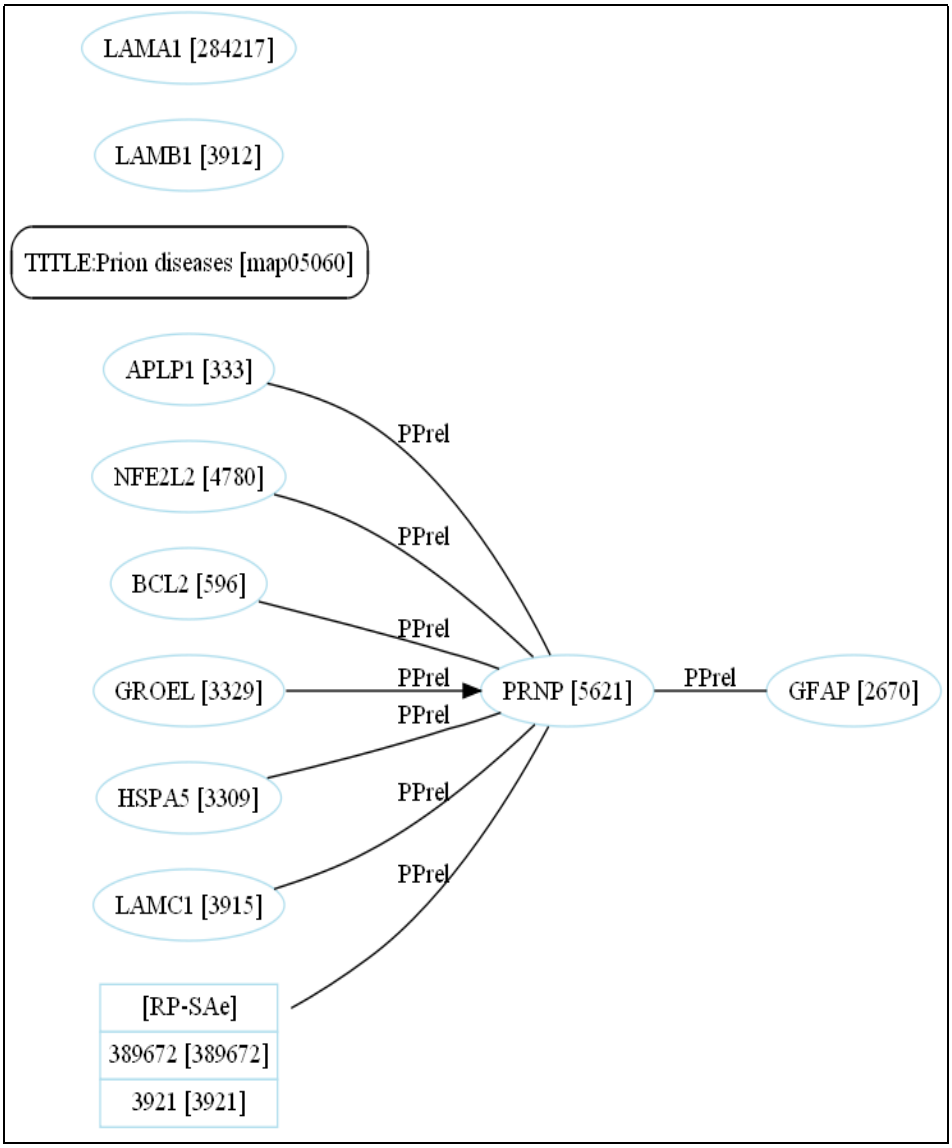
Implementovaný program dává různé formy výsledků. Mezi implementované funkce patří: informace o dráze, porovnání dvou drah, porovnání drah uvnitř dvou (jednoho) adresáře a shlukování drah nad adresářem. Následující text představí tyto funkce a uvede příklady výstupů programu.

Funkce informace o dráze vypíše na standardní výstup: jméno a číslo dráhy, NCBI číslo a oficiální názvy genů obsažených v dráze. Pokud není oficiální jméno v XML souboru uvedeno, je vypsáno opět číslo NCBI. Následně se program pokusí ověřit nalezené geny v dráze s referenčním souborem (poskytuje ho jen KEGG) a případně vypíše neshody. Nakonec je na výstup vypsán seznam všech složek, které se v dráze vyskytují. Do souboru, který je implicitně pojmenován *draha.dot*, je vygenerován textový popis grafu dráhy určený pro zpracování programem Graphviz.

```
Draha: Prion diseases
Cislo: hsa05060
------------------------------------------------------------------------
Obsahuje geny:
2670    GFAP
284217  LAMA1
3309    HSPA5
3329    GROEL
333     APLP1
389672  389672
3912    LAMB1
3915    LAMC1
3921    3921
4780    NFE2L2
5621    PRNP
596     BCL2
Celkem: 12
------------------------------------------------------------------------
Overeni genu v draze:
Pocty genu nesedi, referencni data: 14 a data z xml: 12
V draze z xml chybi: 3569, 7124,
------------------------------------------------------------------------
Vsechny slozky z drahy:
2670    GFAP
284217  LAMA1
3309    HSPA5
3329    GROEL
333     APLP1
```

```
389672  389672
3912    LAMB1
3915    LAMC1
3921    3921
4780    NFE2L2
5621    PRNP
596     BCL2
Celkem: 12
--------------------------------------------------------------------------
```

Tabulka 4.1: Příklad výpisu informací o dráze Prion diseases z databáze KEGG



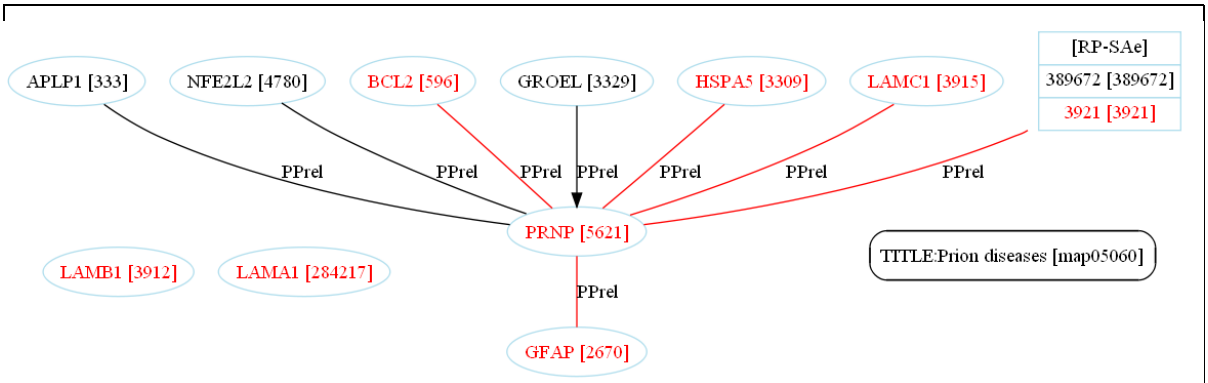
Obrázek 4.1.: Příklad vygenerovaného grafu pomocí nástroje GraphViz.

V hranatých závorkách jsou uváděny NCBI čísla genů, případně jiné identifikační údaje uzlu.

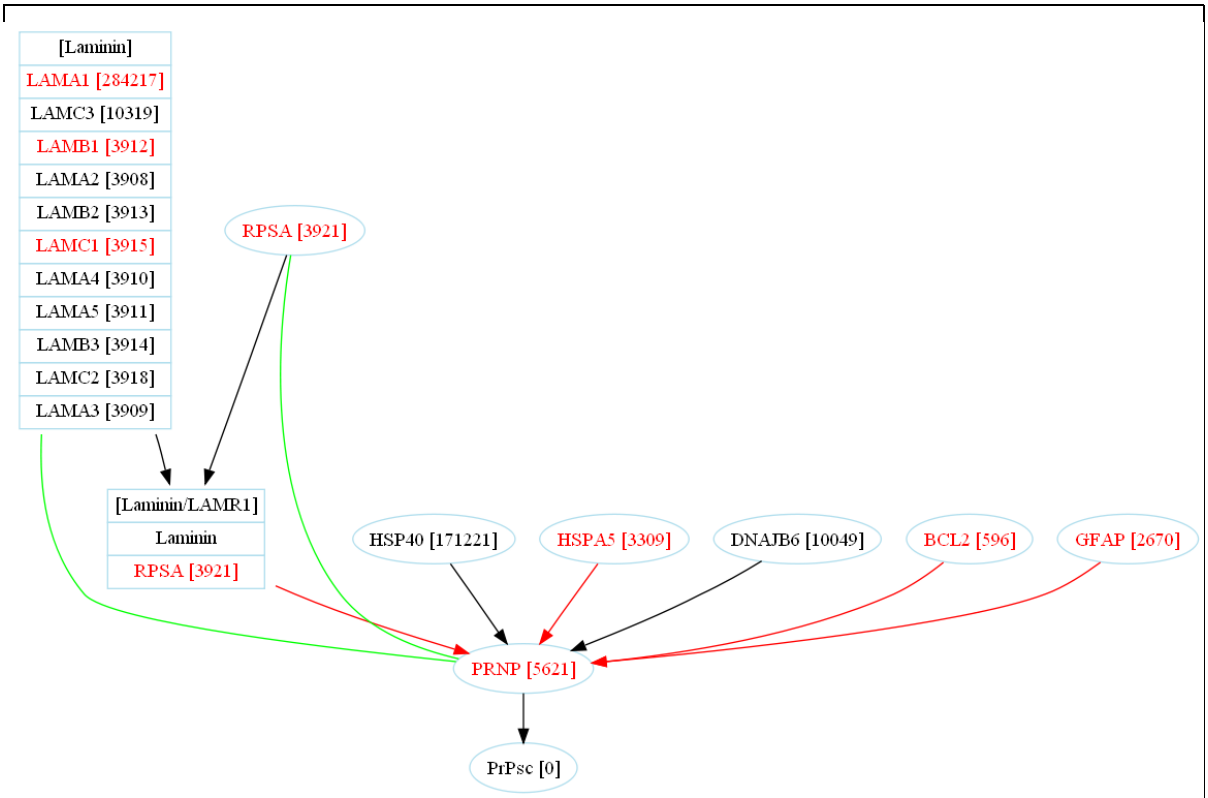
Funkce porovnání dvou drah vypíše na standardní výstup: názvy a čísla srovnávaných drah. Nalezené geny v dráhách jsou, pokud existují referenční údaje, opět překontrolovány, a případně vypsány odlišnosti. Dále je na výstup vypsáno skóre editační vzdálenosti. Je to číslo od 0 do 1, kdy 1 značí velmi blízkou podobnost. Jako další údaj jsou vytisknuty informace o počtu shodných hran mezi společnými geny. Pokud program našel nějakou tranzitivní hranu, je uveden jejich počet. Následuje seznam společných genů. Výpis je zakončen informací o počtu všech shodných složek v drahách (včetně genů). Do adresáře se spustitelným programem jsou vygenerovány dva soubory, které textově popisují grafy drah. Po aplikaci nástroje GraphViz na tyto soubory jsou vytvořeny obrázky. Do nich jsou vyznačeny shodné geny (červeně), hrany (červeně) a tranzitivní hrany (zeleně).

```
Draha1: Prion diseases, hsa05060
Draha2: prion pathway, 100062
------------------------------------------------------------------------
Overeni genu v draze: Prion diseases
Pocty genu nesedi, referencni data: 14 a data z xml: 12
V draze z xml chybi: 3569, 7124,
------------------------------------------------------------------------
Overeni genu v draze: prion pathway
Nelze overit pocty genu v draze, protoze k ni neni reference.
------------------------------------------------------------------------
Score editacni vzdalenosti: 0.5
------------------------------------------------------------------------
Porovnani vazeb mezi shodnymi geny drah:
Drahy maji spolecnych: 7 vazeb.
Z toho 2 hran tranzitivnich.
------------------------------------------------------------------------
Vypis spolecnych genu:
2670    GFAP
284217  LAMA1
3309    HSPA5
3912    LAMB1
3915    LAMC1
3921    RPSA
5621    PRNP
596     BCL2
Celkem: 8
------------------------------------------------------------------------
Porovnani vsech slozek drah:
Drahy maji spolecnych 8 ze vsech porovnatelnych slozek drah.
------------------------------------------------------------------------
```

Tabulka 4.2: Příklad výpisu po porovnání dvou drah



Obrázek 4.2: Výsledek po porovnání drah hsa05060 (KEGG) a 100062 (Biocarta), na obrázku je dráha hsa05060.



Obrázek 4.3: Výsledek po porovnání drah hsa05060 (KEGG) a 100062 (Biocarta), na obrázku je dráha 100062

Funkce porovnání drah uvnitř jednoho adresáře porovná navzájem všechny dráhy. Pokud jsou uvedeny dva adresáře, pak každá dráha z prvního adresáře je porovnána se všemi v druhém adresáři. Výsledkem porovnání je skóre. Dráhy jsou seřazeny sestupně od největší shody. Na standardní výstup jsou vytisknuty informace: celkové skóre, názvy, čísla drah a dílčí skóre při porovnání genů, vazeb, editační vzdálenosti a složek. Údaje jsou odděleny středníky.

Př.: skóre ; dráha 1 ; číslo dráhy 1 ; dráha 2 ; číslo dráhy 2 ; geny ; vazby ; editační vz. ; složky

0.531373 ; prion pathway ; 100062 ; Prion diseases ; hsa05060 ; 0.266667 ; 0.164706 ; 0.1 ; 0

0.080000 ; prion pathway ; 100062 ; phospholipase c-epsilon pathway ; 100070 ; 0 ; 0 ; 0.08 ; 0

0.034921 ; rna polymerase iii transcription ; 100039 ; phospholipase c-epsilon pathway ; 100070 ; 0 ; 0 ; 0.0349206 ; 0

0.025000 ; how does salmonella hijack a cell ; 100037 ; phospholipase c-epsilon pathway ; 100070 ; 0 ; 0 ; 0.025 ; 0

0.019355 ; proteasome complex ; 100061 ; prion pathway ; 100062 ; 0 ; 0 ; 0.0193548 ; 0

0.015385 ; how does salmonella hijack a cell ; 100037 ; rna polymerase iii transcription ; 100039 ; 0 ; 0 ; 0.0153846 ; 0

0.014608 ; proteasome complex ; 100061 ; phospholipase c-epsilon pathway ; 100070 ; 0 ; 0 ; 0.00408163 ; 0.0105263

0.012500 ; proteasome complex ; 100061 ; Prion diseases ; hsa05060 ; 0 ; 0 ; 0.0125 ; 0

0.000000 ; rna polymerase iii transcription ; 100039 ; proteasome complex ; 100061 ; 0 ; 0 ; 0 ; 0

0.000000 ; rna polymerase iii transcription ; 100039 ; prion pathway ; 100062 ; 0 ; 0 ; 0 ; 0

0.000000 ; rna polymerase iii transcription ; 100039 ; Prion diseases ; hsa05060 ; 0 ; 0 ; 0 ; 0

0.000000 ; phospholipase c-epsilon pathway ; 100070 ; Prion diseases ; hsa05060 ; 0 ; 0 ; 0 ; 0

0.000000 ; how does salmonella hijack a cell ; 100037 ; proteasome complex ; 100061 ; 0 ; 0 ; 0 ; 0

0.000000 ; how does salmonella hijack a cell ; 100037 ; prion pathway ; 100062 ; 0 ; 0 ; 0 ; 0

0.000000 ; how does salmonella hijack a cell ; 100037 ; Prion diseases ; hsa05060 ; 0 ; 0 ; 0 ; 0

Tabulka 4.3: Příklad výpisu srovnání drah uvnitř adresáře

Funkce shlukování nad adresářem přebírá od uživatele parametr, který značí hranici míry podobnosti, kdy lze shluky prohlásit za podobné a sloučit je. Je to číslo od 0.1 do 100.0. Shlukování s menší mírou podobnosti vede k tomu, že s největší pravděpodobností bude vytvořen jeden shluk a dráhy uvnitř něj si ve většině případů nebudou vůbec podobné. Na standardní výstup jsou vypsány dráhy, které se shlukování účastní a shluky obsahující jednotlivé dráhy. Shluky mají čísla 0 až n, kde n je počet shlukovaných drah. Některá čísla shluků mohou ve výpisu chybět, protože zanikla v průběhu zpracování.

Při shlukování nad databázemi jsem dospěl k závěru, že se vytvoří několik velkých shluků a menší množství malých. Velké shluky mi nepřišly zajímavé, protože většinou obsahují geny, jejichž

výskyt je v databázi drah častý. Na druhou stranu malé shluky obsahují často geny, jejichž výskyt je v databázi drah vzácný, ne-li jedinečný.

```
Provedu shlukovani nad drahami: 100037, 100039, 100061, 100062,
100070, hsa05060,

Vysledek shlukovani s mirou podobnosti 30%:

0: 100037,

1: 100039,

2: 100061,

4: 100070,

5: hsa05060, 100062,
```

Text 4.4: Příklad shlukování, míra podobnosti 30%

4.1 Program Graphviz

Graphviz je software s otevřeným zdrojovým kódem pro vykreslování grafů. Má několik hlavních programů pro vytváření grafů:

- dot - nástroj ovládaný z příkazového řádku sloužící k vytváření orientovaných grafů
- neato - doplněk k dot pro vytváření neorientovaných grafů
- twopi - vytváří hvězdicové rozvržení grafů
- circo - kruhové rozvržení
- fdp - další rozvržení pro neorientované grafy
- dotty - nástroj s grafickým uživatelským rozhraním k prohlížení a editaci grafů přímo z textového popisu.
- lefty - programovatelný "widget", který zobrazuje grafy zapsané v DOT jazyce a uživateli umožňuje manipulaci s entitami grafu pomocí myši a klávesnice [9].

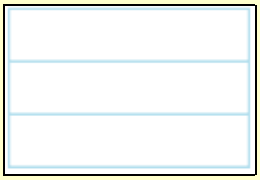
Tyto programy berou jako svůj vstup popis grafu v jednoduchém jazyku DOT a vytváří grafy v několika dobře použitelných formátech (obrázky, SVG pro web, Poscript pro vložení do PDF). Podrobnosti a přehled gramatiky jazyka DOT lze nalézt na [10]. V praxi jsou grafy generovány z externího datového zdroje, ale mohou být vytvářeny a editovány v textovém editoru. Existuje několik programů, které Graphviz používají ke generování grafů např.: Doxygen, GRAMPS, UMLGraph. Dalším užitečným nástrojem je WebDot, cože je CGI program vytvářející grafy z .dot souborů, které pak mohou být vloženy do webových stránek.

Ve své práci jsem použil program Graphviz ke generování obrázků signálních drah. Z velkého množství tvarů hran a uzlů jsem použil následující:

základní uzel(pro geny a složky kromě "compound")

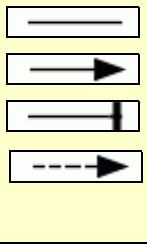
"compound"

makroproces - vícekroková událost zachycená v jednom uzlu, často zprostředkovaná biomolekulou



komplexní uzel složený z více genů, na prvním řádku je název celé skupiny, pak následují jednotlivé složky

"map" - odkaz na jinou dráhu nebo titulek (KEGG)



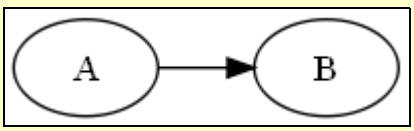
neorientovaná hrana

orientovaná hrana

orientovaná inhibiční hrana

orientovaná hrana typu "indirect effect" nebo "maplink" (propojení s další dráhou)

Vytvořit graf pomocí tohoto nástroje není složité. Velkou výhodu oproti běžnému kreslení v grafickém editoru je automatické generování z textového popisu. S tím souvisí i jednodušší editace. Není tedy nutné při vložení nového uzlu graf ručně poopravovat jako v běžných editorech. Uživatel ovšem za toto pohodlí zaplatí ne vždy pěknými grafy. S tímto problémem je už počítáno a tak Graphviz nabízí program *dotty*. Graphviz také umožňuje pro tvorbu popisků hran a uzlů použít HTML. To, s jakou elegancí lze vytvořit graf, naznačuje následující příklad.




```
echo "digraph Graf { rankdir=LR; A -> B}" | dot -Tpng >test.png
```

5 Závěr

Nejpravděpodobnější překryvy, které jsem při testování programu nad databázemi identifikoval, jsou uvedeny v příloze. S těmito výsledky doporučuji dále pracovat, např. zkusit opětovné porovnání jen těchto dvou drah s vytisknutím grafů pro lepší pohled na společné vazby a geny.

K dosažení lepších výsledků hledání překryvů by přispěl stejný formát poskytovaných dat od všech databází. V práci se sice snažím jednotně reprezentovat dráhu (ať už pochází z jakékoli databáze) jako graf, ale v určitých momentech (např. u vazeb) si nejsem jistý, co můžu považovat za stejné informace, proto prozatím některé vazby zůstávají neporovnatelné.

Z hlediska dalšího vývoje by bylo vhodné systém rozšířit o implementaci knihovní funkce na zpracování další databáze signálních drah. Pro pohodlnější používání je možné vytvořit grafické uživatelské rozhraní. Vývoj systému může také pokračovat v oblasti stanovení skóre, které pak určuje umístění výsledků. Ve vlastní reprezentaci dráhy jako grafu jsou už do budoucna nachystány prostředky pro udělení váhy hranám a uzlům.

Během zpracovávání tohoto projektu jsem se hlouběji seznámil se signální dráhami, s formáty databází KEGG a BioCarta. Vlastní přínos shledávám v seznámení se s využitím informačních technologií ke zpracování biologických dat.

Literatura

[1] Kodíček, M., Biochemické pojmy: výkladový slovník [online], VŠCHT Praha, 2007. Dostupný na URL: http://vydavatelstvi.vscht.cz/knihy/uid_es-002/ebook.html?p=drahy_signalni (květen 2008)

[2] Masopust, J. Patobiochemie buňky. Praha, 2003 Dostupný na URL: https://www.zdravcentra.cz/cps/rde/xchg/zc/xsl/3141_1457.html (květen 2008)

[3] Kanehisa, M., Ontologies and the KEGG, Kyoto University, 2004, s. 6. Dostupný na URL: http://www.sofg.org/meetings/sofg2004/Kanehisa.pdf (květen 2008)

[4] Bělohlávek, R., Vychodil, V. Diskrétní matematika pro informatiky II, Olomouc, 2004, s. 70 Dostupný na URL: http://www.inf.upol.cz/download/study/materials/Diskrétní matematika 2.pdf (květen 2008)

[5] Berghen, F., Small, simple, cross-platform, free and fast C++ XML Parser, 2008. Dostupný na URL: http://www.applied-mathematics.net/tools/xmlParser.html (květen 2008)

[6] Levenshtein distance. In Wikipedia, The Free Encyclopedia. 2008, Dostupný na URL: http://en.wikipedia.org/w/index.php?title=Levenshtein_distance&oldid=199914943 (květen 2008)

[7] National Center for Biotechnology Information, U.S. National Library of Medicine, 2008 Dostupný na URL: http://www.ncbi.nlm.nih.gov/ (květen 2008)

[8] Lorenc, L.,Získávání znalostí z databází: Shlukování, 2004 Dostupný na URL: http://www.fit.vutbr.cz/study/courses/ZZD/public/seminar0304/Shlukovani2.pdf (květen 2008)

[9] Graphviz. In Wikipedia, The Free Encyclopedia. 2008, Dostupný na URL: http://en.wikipedia.org/w/index.php?title=Graphviz&oldid=195753963 (květen 2008)

[10] Graph Visualization Software, Documentation. Dostupný na URL: http://www.graphviz.org/ (květen 2008)

[11] Oficiální stránky KEGG, Kanehisa Laboratories, 2008 Dostupný na URL: http://www.genome.jp/kegg/ (květen 2008)

[12] Oficiální stránky BioCarta, BioCarta, 2008 Dostupný na URL: http://www.biocarta.com/ (květen 2008)

Seznam příloh

Příloha 1. Ovládání programu

Popis ovládání programu a příkladu spuštění. Zadávané parametry musí být v tom pořadí, jak jsou uvedeny. Nutnou podmínkou pro zdárný chod programu je, aby v adresáři, kde je spustitelný soubor programu, byly soubory: `hsa_pathway.list`, `hsa_enzyme.list` a `hsa_ko.list`.

K následnému vykreslení grafů je potřebný program GraphViz. Pro vytváření grafů jsou pod OS Windows nachystány soubory s příkazy, které automaticky vygenerují obrázek.

Popis parametrů:

`-h` vypíše nápovědu

`-i <adresář> <číslo_dráhy>`

vypíše informace o signální dráze a vygeneruje její zdrojový soubor pro graf, <adresář> - název adresáře, kde je dráha umístěna, <číslo_dráhy> - v případě KEGGu je to soubor s dráhou, v případě Biocarty je to Pathway id, př.: `-i biocarta 100062`

př.: `-i kegg hsa05060`

`-c <adresář r> <číslo_dráhy1> <adresář> <číslo_dráhy2>`

porovná dvě dráhy, výsledky vytiskne na standardní výstup, vygeneruje zdrojové soubory grafů obou drah s vyznačenými shodami genů a vazeb.

`-c <adresář1> [<adresář2>]`

provede vzájemné srovnáni drah z <adresář1> příp.: i <adresář2>, pokud je uveden. Porovnání je ohodnoceno skóre, na standardní výstup jsou vytisknuty výsledky.

`-s <adresář> <míra_podobnosti_v_procentech>`

provede shlukování nad všemi dráhami v adresáři, <míra_podobnosti_v_procentech> určuje hranici, při které je možné dráhy prohlásit za podobné, číslo od 0.1 do 100.0, výsledky jsou vypsány na standardní výstup.

Příloha 2. Nejpravděpodobnější překryvy

0.575808 ; toll-like receptor pathway ; 100013 ; Toll-like receptor signaling pathway ; hsa04620 ; 0.20438 ; 0.2 ; 0.171429 ; 0

0.532003 ; t cell receptor signaling pathway ; 100022 ; T cell receptor signaling pathway ; hsa04660 ; 0.162162 ; 0.155556 ; 0.2 ; 0.0142857

0.531373 ; prion pathway ; 100062 ; Prion diseases ; hsa05060 ; 0.266667 ; 0.164706 ; 0.1 ; 0

0.527978 ; egf signaling pathway ; 100181 ; ErbB signaling pathway ; hsa04012 ; 0.140187 ; 0.2 ; 0.162791 ; 0.025

0.493029 ; fc epsilon receptor i signaling in mast cells ; 100165 ; Fc epsilon RI signaling pathway ; hsa04664 ; 0.221154 ; 0.166234 ; 0.0923077 ; 0.0133333

0.488227 ; mtor signaling pathway ; 100101 ; mTOR signaling pathway ; hsa04150 ; 0.173333 ; 0.114894 ; 0.2 ; 0

0.481159 ; proteasome complex ; 100061 ; Proteasome ; hsa03050 ; 0.347826 ; 0 ; 0.133333 ; 0

0.478322 ; tgf beta signaling pathway ; 100017 ; TGF-beta signaling pathway ; hsa04350 ; 0.109091 ; 0.2 ; 0.169231 ; 0

0.476274 ; wnt signaling pathway ; 100002 ; Wnt signaling pathway ; hsa04310 ; 0.113636 ; 0.162637 ; 0.2 ; 0

0.472906 ; pdgf signaling pathway ; 100077 ; ErbB signaling pathway ; hsa04012 ; 0.141593 ; 0.145455 ; 0.163636 ; 0.0222222

0.466657 ; tpo signaling pathway ; 100012 ; ErbB signaling pathway ; hsa04012 ; 0.144144 ; 0.1375 ; 0.162791 ; 0.0222222

0.462300 ; igf-1 signaling pathway ; 100136 ; ErbB signaling pathway ; hsa04012 ; 0.11215 ; 0.194595 ; 0.155556 ; 0

0.460403 ; insulin signaling pathway ; 100124 ; Insulin signaling pathway ; hsa04910 ; 0.0604027 ; 0.2 ; 0.2 ; 0

0.456696 ; antigen processing and presentation ; 100107 ; Antigen processing and presentation ; hsa04612 ; 0.0693069 ; 0.193103 ; 0.194286 ; 0

0.452495 ; bcr signaling pathway ; 100227 ; Fc epsilon RI signaling pathway ; hsa04664 ; 0.150943 ; 0.183333 ; 0.107692 ; 0.0105263

0.451896 ; bcr signaling pathway ; 100227 ; VEGF signaling pathway ; hsa04370 ; 0.127451 ; 0.147368 ; 0.162791 ; 0.0142857

0.449890 ; mapkinase signaling pathway ; 100113 ; GnRH signaling pathway ; hsa04912 ; 0.156863 ; 0.158333 ; 0.134694 ; 0

0.448571 ; ceramide signaling pathway ; 100206 ; Adipocytokine signaling pathway ; hsa04920 ; 0.0952381 ; 0.2 ; 0.133333 ; 0.02

0.436237 ; nf-kb signaling pathway ; 100097 ; Toll-like receptor signaling pathway ; hsa04620 ; 0.121951 ; 0.2 ; 0.114286 ; 0

0.434602 ; pdgf signaling pathway ; 100077 ; VEGF signaling pathway ; hsa04370 ; 0.0927835 ; 0.16 ; 0.163636 ; 0.0181818

0.432479 ; igf-1 signaling pathway ; 100136 ; GnRH signaling pathway ; hsa04912 ; 0.0769231 ; 0.2 ; 0.155556 ; 0

0.431435 ; bcr signaling pathway ; 100227 ; ErbB signaling pathway ; hsa04012 ; 0.118644 ; 0.133333 ; 0.162791 ; 0.0166667

0.428194 ; bioactive peptide induced signaling pathway ; 100226 ; Fc epsilon RI signaling pathway ; hsa04664 ; 0.12037 ; 0.2 ; 0.0972973 ; 0.0105263

0.426285 ; pdgf signaling pathway ; 100077 ; GnRH signaling pathway ; hsa04912 ; 0.105691 ; 0.138776 ; 0.163636 ; 0.0181818

0.425709 ; tpo signaling pathway ; 100012 ; VEGF signaling pathway ; hsa04370 ; 0.0947368 ; 0.15 ; 0.162791 ; 0.0181818

0.424560 ; pdgf signaling pathway ; 100077 ; Fc epsilon RI signaling pathway ; hsa04664 ; 0.128713 ; 0.173913 ; 0.109434 ; 0.0125

0.422125 ; bcr signaling pathway ; 100227 ; B cell receptor signaling pathway ; hsa04662 ; 0.168421 ; 0.133333 ; 0.103704 ; 0.0166667

0.421614 ; akt signaling pathway ; 100245 ; mTOR signaling pathway ; hsa04150 ; 0.0588235 ; 0.2 ; 0.162791 ; 0

0.421490 ; igf-1 signaling pathway ; 100136 ; VEGF signaling pathway ; hsa04370 ; 0.0659341 ; 0.2 ; 0.155556 ; 0

0.419129 ; ras signaling pathway ; 100047 ; mTOR signaling pathway ; hsa04150 ; 0.056338 ; 0.2 ; 0.162791 ; 0

0.418996 ; tpo signaling pathway ; 100012 ; Fc epsilon RI signaling pathway ; hsa04664 ; 0.111111 ; 0.18 ; 0.115385 ; 0.0125

0.417616 ; igf-1 signaling pathway ; 100136 ; Insulin signaling pathway ; hsa04910 ; 0.0759494 ; 0.2 ; 0.141667 ; 0

0.415682 ; erk1/erk2 mapk signaling pathway ; 100170 ; ErbB signaling pathway ; hsa04012 ; 0.110092 ; 0.191304 ; 0.114286 ; 0

0.412767 ; bcr signaling pathway ; 100227 ; GnRH signaling pathway ; hsa04912 ; 0.109375 ; 0.126316 ; 0.162791 ; 0.0142857

0.409385 ; vegf hypoxia and angiogenesis ; 100006 ; VEGF signaling pathway ; hsa04370 ; 0.128713 ; 0.2 ; 0.0571429 ; 0.0235294

0.405848 ; igf-1 signaling pathway ; 100136 ; Fc epsilon RI signaling pathway ; hsa04664 ; 0.0947368 ; 0.2 ; 0.111111 ; 0

0.405379 ; egf signaling pathway ; 100181 ; GnRH signaling pathway ; hsa04912 ; 0.0940171 ; 0.128571 ; 0.162791 ; 0.02

0.401650 ; insulin signaling pathway ; 100124 ; ErbB signaling pathway ; hsa04012 ; 0.0612245 ; 0.2 ; 0.140426 ; 0

0.401548 ; tpo signaling pathway ; 100012 ; GnRH signaling pathway ; hsa04912 ; 0.0826446 ; 0.137931 ; 0.162791 ; 0.0181818

0.400824 ; bioactive peptide induced signaling pathway ; 100226 ; VEGF signaling pathway ; hsa04370 ; 0.0865385 ; 0.2 ; 0.1 ; 0.0142857

0.399825 ; g-secretase mediated erbb4 signaling pathway ; 100172 ; ErbB signaling pathway ; hsa04012 ; 0.0631579 ; 0.186667 ; 0.15 ; 0

0.398989 ; erk1/erk2 mapk signaling pathway ; 100170 ; MAPK signaling pathway ; hsa04010 ; 0.0573477 ; 0.170213 ; 0.171429 ; 0